

CHAPTER 3: RESULTS OF THE MARCH 2001 ADMINISTRATION

Introduction

The legislation establishing CAHSEE called for the first operational form(s) of the exam to be administered in Spring 2001 to 9th graders in the Class of 2004. At the first administration, 9th graders could volunteer, but were not required, to take both portions of the exam. Students who did not pass the exam in that administration would be required to take the exam as 10th graders in Spring 2002.

In Fall 2000, the Superintendent set testing dates of March 7, 2001 for the English-language arts (ELA) portion of the CAHSEE and March 13 for the Math portion. Additional testing dates were set in May (May 17 for ELA and May 24 for Math) for year-round schools that were not in session during the March testing dates. Since participation was to be voluntary, no provision was made for makeup sessions for students who were absent on the designated testing dates.

At the December meeting of the State Board of Education (SBE), the Secretary of Education announced that urgency legislation was being introduced in the state legislature that would change the nature of the first administration. Specifically, the March 2001 administration would be changed to a practice test, introducing 9th graders in the Class of 2004 to the nature and format of the examination, but not classifying any students as either passing or failing the exam. The first operational administration would be in Spring 2002, when all 10th graders in the Class of 2004 would be required to participate. The change was motivated by two concerns. First, it appeared that many students do not complete courses that cover the content of the exam until the 10th grade. Making the test operational for 9th graders could raise significant questions about inequity in opportunity to learn the material covered by the test.

The second reason for the change was that census testing of 10th graders in 2002 would provide important normative information. Under the original plan, no single administration would include a representative sample of students. The Spring 2001 administration would be voluntary and the Spring 2002 administration would partially or completely exclude students who had previously passed one or both parts of the exam. Before operational results could be reported, the Board had to determine the minimum score levels required for passing each of the two parts. Minimum passing scores based on performance results on previous administrations of a test are often referred to as “performance standards,” in contrast to content standards, which describe the material covered by the test. In setting performance standards, it is common for the governing body to use normative information (specifically the proportion who pass the exam) to check on the reasonableness of performance standards recommended by panels of content experts.

Following the December 2000 Board meeting, Senate Bill 84 (SB 84) was introduced to enact changes with respect to the initial administration of the CAHSEE. SB 84 was introduced in the state Senate on January 11, 2001 as an urgency measure, meaning that it would take effect immediately. Otherwise the bill would not become effective until well after the planned March administration of the test. The Senate Education Committee approved the

bill with amendments on February 1, 2001 and by the full Senate on February 20, 2001. In the Assembly, the bill was amended to return it to its original form, deleting the Senate amendments that included a provision to defer the requirement until the Class of 2005. On March 1, 2001, the Assembly passed the bill in its original form. Assembly amendments restored the urgency provision, which had been deleted in the Senate. As an urgency measure, the bill required approval by 60% of the members of each house. When the Senate voted on the revised (original) measure on March 1, 2001, it failed to receive the required 60% majority. A second vote was taken on March 5, 2001, but the bill again failed to obtain the required majority. Note that the final vote to defeat SB-84 occurred just 2 days before the administration of the ELA portion of the exam, scheduled for March 7. Fortunately, most 9th graders were already signed up to take the exam, but it is likely that many would have received more extensive preparation had it been known earlier that the exam would count. In reality, however, students in the Class of 2004 were not negatively impacted by the failure of the legislation. They now had one more chance to pass the exam, which they would not have had if it had gone through.

In the remainder of this chapter, we describe the initial administration of the CAHSEE in Spring 2001 and discuss our analysis to date of results from this administration. Since data from the May administration are not yet available and final decisions about scoring and reporting rules are just being made, our analysis as of the end of Year 2 of the evaluation is necessarily preliminary. During Year 3, we will complete analysis of the results and submit a report to the Superintendent, State Board, Governor, and legislature by February 1, 2002 as required under EC60855.

Administering CAHSEE

The plan for administration of a practice test in Spring 2001 would also have allowed an opportunity for a dry run of test administration procedures. As described below, the joint demands of fairness and test security placed a number of difficult constraints on the administration of the CAHSEE. These constraints impacted schools and districts differently, depending on the number of students to be tested, how student time is normally scheduled, the availability of testing space, and other factors. In this section, we describe our observations of the Spring 2001 administration and offer some suggestions for consideration in future administrations of the CAHSEE.

Sources of Information

HumRRO collected information on test administration of CAHSEE from three sources:

- Observing three schools as they administered CAHSEE
- Monitoring training workshops for school and district personnel responsible for test coordination before the March administration and a focus group of district test coordinators after the March administration
- Surveying a modest sample of school test coordinators

Characteristics of the test sessions observed are shown in Table 3.1. The HumRRO observer watched students take the test—attending to the pace of progress, test security, and level of distraction—and interviewed the test coordinators. While the schools varied in the ways they conducted CAHSEE, school staffs were well-prepared and provided good test conditions. The most striking overall feature was how seriously the students took the test.

TABLE 3.1 Characteristics of Schools Observed

School	Subject	School Type	Approximate Number Tested	Environment	Accommodations
A	ELA (March)	Urban	850	Classrooms	None
B	Math (March)	Rural	275	Auditorium	None
C	ELA (May)	Suburban	575	Classrooms	Special Education (Separation)

Our Spring 2001 survey of teachers and principals in the longitudinal sample of schools we are following included a brief survey of site coordinators. The site-coordinator survey (see Appendix C) asked for feedback on guidance received, students tested, the general approach to conducting the test, and changes planned for future administrations of CAHSEE. Coordinators for 42 schools returned the survey. About half had the title of test coordinator and another third were assistant principals.

CDE conducted a focus group with about 40 district testing coordinators between the March and May test dates to collect feedback on test logistics. The coordinators rotated through four stations to discuss issues with administering CAHSEE: (a) testing manuals, workshops, and staff development; (b) logistics, scheduling, and security; (c) test administration support; and (d) accommodations and regulations. The discussion of results from all three sources is organized by those topics.

Observations on Test Administration

Testing Manuals, Workshops, and Staff Development

The test developer and its subcontractor for processing and reporting (NCS Pearson) conducted five workshops with district and school test coordinators (HumRRO observed one of the workshops). The theme of the workshops was that CAHSEE was important and the coordinators needed to get immersed quickly and take seriously the administration of the tests. Topics included session length, test security, and score reports. Speakers walked coordinators through the “aggressive” requirements to receive materials, prepare answer documents, and return materials.

About 60% of the surveyed coordinators had read at least one of the coordinator manuals, but only half reported reading Directions for Administration. Most thought that the information in the manuals was clear, but several suggested changes, including: (a) Combine the coordinator manuals to eliminate overlap, (b) reduce restrictions on distribution of directions for administration, and (c) clarify the instructions for filling out the answer documents.

About 25% of the school site coordinators in the survey had attended the workshop. Although they generally felt frustrated by the uncertainties of whether the test would count,

the only negative comment about the content of the workshop was that not enough of it was about logistics, especially what to do with students who were not being tested. One response to a question about plans for the next administration was, “Going to the conference was extremely helpful. Other site coordinators from my district did not go and they were confused. I recommended to them that they go to the meeting next time!”

While coordinators who attended the focus group also thought that the Directions for Administration were confusing, especially regarding the completion of background information if the school had taken advantage of the precode option, they were positive about the workshops. They said that the workshops should be conducted earlier, at more sites, and with fewer people per session.

CDE supported staff development through presenter workshops and teacher guides. Comments from the focus group about those efforts were strongly positive, especially for the option to access information via the Internet.

Logistics, Scheduling, and Security

Feedback in this area concerned extended test-taking time, breaks, the length of the ELA test, and options for other students.

The main logistics problem in the observed schools was balancing the option of extended time for students who needed it with test security and test conditions. School A did not provide extended time but had very good test security. At the end of both sessions, proctors alerted students that time was almost up and they should finish the test; they did not mention that additional time was available. Everyone took a break between sessions. Because this school allotted over 2 hours for each session, all students appeared to finish by the scheduled time, but some students in each session clearly rushed to complete their essays.

School B provided extended time and preserved testing conditions but did so at the cost of test security. This school tested students in an auditorium with lapboards and allowed about 3 hours for testing (because they did not precode answer documents, completing the background section took 30 minutes). Students ignored the section breaks, moving directly to Section 2 as soon as they completed Section 1. After students finished Section 2, they left the auditorium. Even though students had a chance to change their answers based on information they got during the break, the approach minimized disruptions for more deliberate students. About 5% of the students had not finished by the time lunch started. They were released to lunch and told to report to a classroom to complete the test. Although this model was not typical of the schools in the survey, it was not unique: Two other schools disregarded the sections (and another plans to next time); five allowed students to finish the first section after the break; and six had students finish after lunch.

School C tested students in classrooms but had not given proctors guidance on extended time because feedback from schools that had tested in March was that time was adequate. As a result proctors gave a variety of options to students who needed more time. In some classes, such students were sent to the library. In another class, students were told they could work through the break but no longer than that. Some students who needed time for Section 2

continued through lunch and received compensatory time for lunch. A survey respondent wrote: “When students need more time, it is a logistical nightmare.”

A consistent comment from all sources was that the ELA exam was too long. For example, a district coordinator commented that “kids max at 2 ½ hr,” and a proctor at an observed school said, “These kids are fried.” As a result of similar comments, CDE has established a schedule for 2002 that will test ELA over 2 days. The length of the math test was not cited as a problem, but district coordinators cautioned that the apparently comfortable time requirements might be because many students who lacked algebra skills did not do those calculations. For math, only about 1% of the students failed to answer the last question on the test. For ELA, approximately 9% of the students did not attempt the final question, which was an essay.

Schools were also concerned about what to do with other students during testing. School A had a schoolwide writing activity, which freed up classrooms and teachers, and gave flexibility for the lunch schedule, but also resulted in significant absenteeism. Two other schools had special schoolwide activities. Focus-group coordinators reported that other schools scheduled field trips and minimum days. Most of the surveyed schools held to a regular class schedule for other students and about 25% conducted regular classes with a revised schedule. Only seven schools reported lower attendance than normal by other grades.

School and district coordinators requested the option of Saturday testing or using non-instructional days for testing. At the coordinator focus group, CDE explained that the Saturday option was impractical because, under the California Education Code schools could not mandate Saturday attendance.

Several school site coordinators from both the observation and survey samples reported concern about logistics for 2002 related to having to test 9th-grade volunteers as well as 10th graders who did not pass in 2001. At the coordinator focus group, CDE said that legislation had been introduced to eliminate testing of 9th graders.

Test Administration Support

Support included the option of precoding identification to answer documents, delivery of materials, and hotline support from AIR and NCS. Comments from all sources were overwhelmingly positive. About 75% of the respondents to our survey reported taking advantage of precoded answer documents, and the same number said they will use the option again. One school coordinator considered CAHSEE the easiest to administer of all statewide tests the school conducts (excluding logistics).

Accommodations and Regulations

Two of the observed schools did not provide any accommodations for English learner (EL) students or students with disabilities. One of those two schools encouraged special education students to opt out of CAHSEE, and the other tested all students without regard to status. The only school that tested special education students at all differently grouped the students with their regular classes in their regular rooms, which allowed the proctor to give special attention to instructions. The special education students did not need extra time; in

fact, their biggest problem seemed to be maintaining effort through the session. After 1 hour, most had finished and all but one had finished after 1 hour and 15 minutes. In contrast, fewer than 10% of students in a regular session were finished after 1 hour, and the modal completion time was about 90 minutes.

Although two of the observed schools had high populations of Spanish speaking students, no school offered the option of using glossaries. In fact, there were no official glossaries for the 2001 administration since the regulations permitting glossaries had not been finalized. There was a place on the answer sheet to indicate that glossaries were provided and apparently some form of glossary was provided to a few students. Similarly, regulations regarding calculators were not yet finalized. There was no place on the answer sheet to indicate that calculators were provided, but seven testing coordinators responding to our survey indicated calculator use.

The surveys also reflected a low frequency of accommodation. School site coordinators reported 16 cases in which special education students took advantage of calculators, glossaries, readers, or large-format materials. Because some district coordinators in the focus group raised the possibility that students in large schools might have more access to accommodations than others, the distribution of accommodations by school size is shown in Table 3.2. Although the number of accommodations is too small for any final conclusion, the number of accommodations offered per school in the sample is virtually the same for small schools (.45) as for large schools (.47)

TABLE 3.2 Accommodation for Students With Disabilities by School Size *

	Enrollment:	501+	100-500	1-99	Total
	Number of Schools:	17	14	11	42
Accommodation					
Calculator		4	0	3	7
Glossary		0	1**	0	1
Reader		3**	2	2	7
Large Format		1	0	0	1

* Based on our Spring 2001 survey of 42 test coordinators in our longitudinal study sample.

** Also for EL (English learners)

Table 3.3 shows the number of students who were provided various accommodations according to information recorded on the student answer sheets. At this time it is not fully clear how different schools interpreted the reporting categories used. *Scheduling* accommodations, for example, generally meant additional breaks, since all students were to be allowed almost unlimited time. This was clearly the most frequent accommodation. *Presentation*, the next most frequent accommodation, generally meant large format text.

Accommodations for EL were even less frequent. As shown in Table 3.2 above, only one school in the survey offered glossaries to EL students and one provided the option of a reader. Coordinators were asked to identify other accommodations. These included separate rooms (two special education; one EL), extended time (three special education), and a bilingual aide (EL).

TABLE 3.3 Accommodations Reported for All Students Testing in March 2001

Accommodation	ELA		Mathematics	
	Number	Percent	Number	Percent
Scheduling	6,712	1.92	6,403	1.85
Presentation	1,530	0.44	880	0.25
Braille	108	0.03	40	0.01
Response	924	0.26	1102	0.32
Glossary	403	0.12	118	0.03
Test Read Aloud	N/A	N/A	1564	0.45

The extent of accommodations was no doubt affected by uncertainty about whether results would count for graduation, which may have led to reduced participation of special education and EL students. About 40% of the surveyed coordinators reported that they tested fewer than half of the eligible students with disabilities and about 30% of EL students. In addition, coordinators in the focus group reported confusion about which means of accommodation were available. Consistent with those reports, about 40% of the school coordinators expected more accommodation in 2002.

Recommendations for Future Test Administrations

Logistics, Scheduling, and Security. The plan to conduct the ELA session over 2 days is a good idea. It will greatly reduce the most severe problems with extended time and test security. However, the problems also apply, on a smaller scale, to math. Coordinators in the focus group requested models of approaches that are effective. At least two models should be developed for math to cover classroom and large-group (i.e., gym or auditorium) environments. Both should have a recommended end-of-session statement that makes the option of additional time explicit and should include arrangements for a room and a trained proctor to provide extended time. When developing the large-group model, it would be desirable to consider more flexibility in security measures.

Test Administration Support. The high-quality support should continue.

Testing Manuals, Workshops, and Staff Development. The only problems were the clarity of one of the manuals, Directions for Administration, and availability of the workshop. CDE has directed the developer to revise the Directions for Administration. It would be a good idea to continue the workshop at least for 2002 with a greater emphasis on increasing the number of school site coordinators who participate. The workshop should include breakout sessions for coordinators who will test in classrooms and those who will conduct large-group sessions.

Accommodations and Regulations. CDE is increasing coordination with Special Education coordinators and advocates. In the new development contract, the Department is also requiring the developer to produce second-language glossaries for the mathematics test, and is seeking legislative clarification on the intent of the EL waiver. These actions, plus stable expectations for 2002 testing, should result in more widespread use of accommodation options and a better sense of whether guidance is adequate. Because the increase in accommodation will require logistical support, we recommend that Special Education coordinators be invited to attend the coordinator workshops, if possible with their test

coordinator. CDE or the test developer should conduct breakout sessions on logistical support for the accommodations.

Review of Item Statistics

We computed item statistics based on all of the roughly 350,000 students taking each of the two exams. Items performed close to original expectations with respect to the difficulty and information value of each item. No significant problems were found.

We selected a random sample of 9,000 students for each exam and used their responses to compute item response theory (IRT) parameter estimates. Item response theory parameters provide a function indicating the probability of a correct response (or particular score level for the essay questions) for students at a given (but unobserved) level of achievement. This function is typically used in reverse to estimate the unobserved level of achievement from the observed patterns of correct and incorrect responses. AIR used a relatively parsimonious IRT model (1-parameter) that leads to a clear relationship between number correct and underlying scale scores. In our analyses, we used more complex models—the 3-parameter logistic model (Lord & Novick, 1968) for multiple-choice questions and an 8-level partial credit model (Muraki, 1992; 1997) for the essay questions. Our purpose in fitting these models was not to develop the reporting scale, but to provide estimates of score accuracy that were as accurate as possible.

In our February 2002 report, we will compare item statistics from the test forms used in the March and May administrations. Data from the May administration was not available at the time this report was written.

Review of Item Scoring Procedures

HumRRO staff observed training of the table leaders and then the individual scorers who rated the responses to each of the two essay questions. Briefly the scoring process worked as follows:

- Each essay was independently scored by two different judges.
- If the judges both agreed that the paper was unscorable or if they both gave scores and these scores did not differ by more than 1 point then the final score was the average of the two judges' ratings (or 0 if they both agreed the response was unscorable). Differences of one point were expected for papers near the boundary of the scoring levels ("fence sitters").
- If the judges disagreed as to whether the response was scorable, or if they gave scores that differed by two or more points, the paper was read and scored by a third scorer (usually the table leader). If the third judge agreed with one of the first two judges, then that rating was the final score.
- It was often the case that the 3rd judge gave a different rating than either of the first two judges, usually a rating falling between the ratings of the first two judges. In this

case, a 4th judge (who was generally more experienced in the scoring process) read the paper. The 4th judge's rating, which always agreed with the ratings of one of the first 3 judges, was taken as the final score for the essay.

Table 3.4 shows the frequency of agreement between the first two judges and the frequency of different ways in which initial disagreements were resolved.

TABLE 3.4 Scoring Agreement for the Essay

Result	First Essay Question		Second Essay Question	
	Frequency	Percent	Frequency	Percent
Absolute Agreement	260,381	74.4%	226,831	64.8%
Difference of 1 Point	85,586	24.5%	115,214	32.9%
Disagreement over Scorability	669	0.2%	508	0.2%
Scorable, but difference > 1	2,202	0.6%	4,182	1.2%

As indicated in the above table, disagreements by 2 points or more were quite rare. The first two judges reached sufficient agreement more than 99% of the time for the first essay and roughly 98% of the time for the second essay. Where disagreements did occur, there was a reasonable process for their resolution.

Setting the Minimum Passing Score

The Score Scale

Efforts to determine the minimum performance required for passing each test focused on a total points, or raw score, scale for the form of each test used in the March 2001 administration. The primary question was how many of the maximum possible raw score points a student must obtain to pass the exam.

At the first stage of scoring, a "raw score" is computed for each student. *For mathematics*, the raw score is simply the number of questions answered correctly. *For ELA*, the raw score is a weighted combination of the number of correct answers to the multiple-choice questions and the student's scores on each of the two essays. The exact equation is:

$$\text{Raw Score} = .7683 * \text{MC} + 3.3750 * \text{CR}$$

Where MC is the number of multiple-choice items (out of 82) answered correctly and CR (constructed response) is the sum of the two essay scores, each of which ranges from 0 to 4 in half-point increments (except that it is not possible to get a score of 0.5). For mathematics, the raw scores range from 0 to 80. For ELA, the maximum possible raw score is $.7683 * 82 + 3.3750 * 8 = 90$. For ELA, the raw scores are rounded to whole numbers.

As with most testing programs, scores ultimately will be reported on a standardized scale. Raw scores are not exactly comparable across test forms due to minor differences in the difficulty and information value of the questions in each test form. Scores on this standardized scale will be comparable across different test forms. A separate translation will be developed for each different test form mapping the raw scores into scale scores. The initial score scale will be a linear translation of the Rasch (one-parameter) IRT scale (see for

example, van der Linden & Hambleton, 1997) developed from the March administration. It is expected to range from 250 to 450 with the passing level somewhere near the middle. Plans for projecting raw scores from subsequent forms (including the test form used in May 2001) have been outlined, but not extensively reviewed.

Standard Setting Panels

The test developer negotiated a subcontract with Howard Mitzel of Pacific Metrics to conduct a standards-setting workshop using the bookmark procedure explained below. The workshop was conducted May 18–20, 2001. Two HumRRO observers attended the workshop.

CDE had arranged for 90 workshop participants, 45 each for ELA and mathematics. Most participants were classroom teachers or content specialists who had been nominated by their districts. In addition, the roster included university faculty, school and district administrators, parents, and business people. About 10 had been on the CAHSEE Panel or Technical Advisory Committee. Almost all panelists participated in all sessions on both days. As a whole, the panels were broadly representative of the state and knowledgeable about the California content standards and high school curriculum. Individually, the level of commitment and effort was high.

The bookmark procedure was appropriate for the purpose and was implemented faithfully. The process began with a general orientation and an opportunity for each participant to take an abbreviated form of the exam. At the orientation, Mitzel stressed the need to make decisions based on test content. He described the ordered-item booklets, one each for mathematics and ELA, which listed the test questions in order of difficulty based on the March administration. For each question, participants were to discuss what made the question more difficult than the preceding questions, with particular attention to other questions from the same content strand.

Participants next moved to rooms for their content area, where they worked in groups (tables) of five or six participants, one of whom had been trained to be a table leader. Each table appeared to follow the directed procedure for discussing the knowledge and skills required by each question. A list showing the specific content standard assessed by each item was given to the math group and several tables noted that there were easy and difficult questions for each of the content strands into which the standards are organized.

After each table had discussed each of the test questions, the entire group reconvened for training on how to place a bookmark. Each participant was to place a marker to divide two item sets: items covering material each student should know and items covering material that is "maybe not needed" to get a diploma. Mitzel emphasized the differences between the bookmark placement and number-correct scores. After the training, participants worked individually to place the marker in their ordered-item booklets.

The next day, each table received a summary of individual bookmarks for the table showing the lowest, highest, and median bookmark placement. Table members discussed the rationale for their initial bookmark placements. Following this discussion, each panelist provided a revised bookmark placement. After lunch, the revised results were presented,

showing the median bookmark and range for each table, along with what the pass rate would be for the median for the room. For math, many, but not all, were surprised by how low the projected pass rates were. The rate for ELA seemed to be what most participants expected. A representative from each table then described the rationale(s) for the table. Most were optimistic about the potential for students to improve during the 10th and possibly 11th grades. The median ratings did not change based on the impact information. One change that might be considered in future workshops would be to report the passing rates associated with the minimum and maximum bookmark placements in addition to reporting the passing rate for the median bookmark placement. This would give participants a better understanding of the level of consensus they had achieved.

In the end, both panels recommended that the minimum passing score be set at 70% of the total possible points on each test. Though that is suspiciously close to traditional passing grades, we heard no evidence either that participants considered any criterion besides content or collaborated between content areas.

The Final Decision

CDE staff reviewed the panel's recommendations and discussed them with the Superintendent. The Superintendent stated that the recommendations of the standards-setting panel should be considered a long-term goal. She recommended that the provisional passing rates for the initial implementation of the CAHSEE be somewhat more lenient. The specific recommendation, 60% of total possible points for ELA and 55% for Math, reflected the fact that the current content standards had not been in place when members of the Class of 2004 were developing prerequisite skills. She also recommended that the State Board of Education should reexamine the test scores after students in the Class of 2004 are well into the 10th grade curriculum to determine whether students are passing in sufficient numbers to demonstrate that adequate opportunities to learn are being provided. On June 7, the Board adopted the passing standards recommended by the Superintendent.

Who Passed?

Once the minimum passing scores were established, it was possible to conduct a number of analyses to see who passed each of the two parts of the exam. A major charge for our evaluation is to report passing rates for specific demographic groups, including all students, economically disadvantaged students, students with disabilities, and EL students. Table 3.5 shows our estimates of the passing rates for each of these groups and also by gender and race. It should be noted that these estimates are based on initial data files supplied by AIR and NCS and do not include results from the May administration. Final counts including the May results will be included in our February 2002 report to the legislature.

The preliminary data files were not merged and did not contain student identifiers that would allow us to see how many students passed both parts of the test. Merged information will be available in August when the scores are issued.

TABLE 3.5 Passing Rates for each Test

Group	Sex	ELA		Mathematics	
		Number	Pct. Pass	Number	Pct Pass
All Students	All	349,938	64.59	345,810	44.65
	Female	171,161	71.52	169,070	43.26
	Male	177,608	58.03	175,304	46.10
African American	All	28,374	50.22	27,930	24.54
	Female	14,272	59.89	14,066	24.51
	Male	14,003	40.42	13,759	24.65
Asian	All	30,373	76.79	30,579	70.75
	Female	14,644	81.52	14,768	70.23
	Male	15,678	72.43	15,746	71.28
Caucasian	All	127,494	81.95	125,293	63.69
	Female	62,442	88.26	61,373	62.37
	Male	64,799	75.88	63,628	64.99
Hispanic	All	140,710	48.68	138,709	25.58
	Female	69,156	56.04	68,172	23.64
	Male	71,224	41.60	70,190	27.50
Economically Disadvantaged	All	108,847	46.18	107,692	25.98
	Female	52,157	53.57	51,654	24.06
	Male	56,524	39.41	55,840	27.79
Students with Disabilities	All	32,421	22.46	31,857	12.33
	Female	11,011	27.54	10,773	9.40
	Male	21,337	19.88	20,940	13.84
English Learners	All	47,621	29.72	47,497	16.93
	Female	22,156	35.33	22,086	14.69
	Male	25,361	24.89	25,276	18.92

The ELA test combined multiple-choice and essay questions. One question that was debated extensively by the CAHSEE Panel was how well students should have to perform on each part in order to be considered proficient. In the end, separate passing levels were not established for each question type or for different content levels. The result was a compensatory model, where exceptional performance in one content area or on one type of question would compensate for lower performance in other content areas or on other types of questions.

Table 3.6 below shows the number of students with each possible total essay score (the sum of the scores on the two essays) and the percent of these students who will receive a passing score on the ELA exam. A very small number of students (242) passed the ELA exam without writing either of the essays. Nearly all of the students who passed ELA (more than 99%) had a total essay score of at least 3.0, meaning that two of the four judges rated one or the other of their essays at score level two or higher. Roughly 94 percent of the students who passed received a total essay score of 4.5 or higher, meaning that they must have received a score of at least 2.5 on one of their two essays. Thus nearly all students who passed the ELA exam received a score of 3 or higher on the 4-point rating scale from at least one of the four judges who rated their essays.

TABLE 3.6 Percent Passing the ELA Exam by Total Essay Score

Total Essay Score	No. of Students	% of Students	No. Passing ELA	% Passing ELA
0.0	15,920	4.5%	242	1.5%
1.0	5,968	1.7%	104	1.7%
1.5	3,100	0.9%	68	2.2%
2.0	12,096	3.5%	753	6.2%
2.5	7,494	2.1%	689	9.2%
3.0	14,693	4.2%	2,369	16.1%
3.5	11,494	3.3%	2,382	20.7%
4.0	24,772	7.1%	7,763	31.3%
4.5	26,077	7.5%	12,410	47.6%
5.0	39,320	11.2%	25,497	64.8%
5.5	43,508	12.4%	34,629	79.6%
6.0	65,278	18.7%	59,761	91.5%
6.5	37,004	10.6%	36,214	97.9%
7.0	24,425	7.0%	24,357	99.7%
7.5	12,253	3.5%	12,248	100.0%
8.0	6,536	1.9%	6,536	100.0%
Total	349,938	100.0%	226,022	64.6%

Table 3.7 shows a similar breakout of passing rates for different number correct scores on the multiple-choice questions. It was not possible to receive a passing total score without answering at least 36 of the multiple-choice questions correctly. The essay score translated to a maximum of 27 of the 90 possible total score points and a score of 54 was required for passing. At least 36 multiple-choice questions had to be answered correctly to achieve a score of 27 on the multiple-choice portion of the ELA exam. In fact, no one passed the exam without answering at least 38 of the 82 multiple-choice questions correctly. Students who answered 71 questions correctly received at least 54 points from the multiple-choice portion and so were guaranteed a passing total score.

TABLE 3.7 Number and Percent of Students Passing the ELA Exam by Total Multiple Choice Score

Multiple Choice Total Score	Number of Students	Percent of Students	Number of Students Passing	Percent Passing for this MC Score
0-37	66,310	18.9%	-	0.0%
38-40	13,269	3.8%	27	0.2%
41-45	24,875	7.1%	2,424	9.7%
46-50	30,156	8.6%	16,639	55.2%
51-55	35,126	10.0%	29,323	83.5%
56-60	40,839	11.7%	38,972	96.2%
61-70	88,495	25.3%	87,769	99.2%
71-82	50,868	14.5%	50,868	100.0%
TOTAL	349,938	100.0%	226,022	64.6%

For mathematics, we examined passing rates for different course completion patterns. Information was recorded on the student answer sheets as to the grade (from 7 to 12) in which specific mathematics courses were taken. Unfortunately, there was no specific way to indicate that a given course was not taken. For 106,987 students, there were no marks for any course in the preliminary data files. The course status of these students was set to missing.

Course status was set to invalid for a few students who indicated courses taken in grades they had not reached. Otherwise, students were classified on the basis of whether they had taken or were taking Algebra 1. Students who took Algebra 1 prior to the 9th grade were further classified according to whether they were or were not currently enrolled in Geometry. Students who had not taken Algebra 1 but had taken or were enrolled in an Integrated Math course were coded separately. Table 3.8 shows the number of students and passing rates for the CAHSEE Math exam for each math course status category. Not surprisingly, students who had completed Algebra 1 and were enrolled in Geometry had a very high passing rate – in excess of 90%. Students who had not taken and were not enrolled in Algebra 1 had very low passing rates – below 20%.

TABLE 3.8 CAHSEE Math Passing Rate by Math Courses Taken

Math Course Status	Number of Students	Percent Passing Mathematics
Completed Algebra and Enrolled in Geometry	35,923	90.29
Completed Algebra, not Enrolled in Geometry	10,819	60.74
Completed or Enrolled in Integrated Math 1	11,283	52.81
Currently Enrolled in Algebra 1	118,097	48.77
Algebra 1 not Taken	61,537	18.23
Course Information Missing	106,987	37.80
Invalid Course Information	1,264	16.67

One key question is the extent of variation in passing rates by school. To the extent that relatively few students from a particular school pass the exam, there is reason to believe that somewhere along the way these students have not had the opportunity to learn either the material covered by the test or, even more likely, key prerequisite skills taught at lower grades. Conversely, if most students in a school do pass the exam, there is good reason to believe that students at that school did have adequate opportunity to learn the required material. Table 3.9 and Table 3.10 below show the number of schools where very few (less than 20%) of the students tested received passing scores through the number of schools where nearly all students (at least 90%) of the students passed. The preliminary data files contained 1,500 different school codes for the ELA exam and 1,501 school codes for the mathematics exam. In nearly a quarter of these schools, fewer than 10 students were tested. For these schools very low or high passing rates are not surprising. Most of the schools where larger numbers of students were tested had passing rates between 25% and 75%, consistent with the overall passing rates for the state as a whole. Schools where at least 100 students were tested and the passing rate was below 25% may deserve special attention.

TABLE 3.9 Number of Schools by Passing Rates and Students Tested – ELA

% Passing in the School	Number of Students Tested				Total Schools
	1-9	10-99	100-400	500+	
0-10%	103	31	1	1	136
10-25%	30	81	10	7	128
25-75%	137	206	234	199	776
75-90%	27	60	148	80	315
90-100%	44	41	50	10	145
Total	341	419	443	297	1500

Note: For schools where 500 or more students were tested, the passing rates ranged from 7.6% to 98.6%; for schools where more than 100 to 499 students were tested, the passing rates ranged from 1.1% to 100%.

TABLE 3.10 Number of Schools by Passing Rates and Students Tested – Mathematics

% Passing in the School	Number of Students Tested				Total Schools
	1-9	10-99	100-400	500+	
0-10%	206	140	13	6	365
10-25%	43	83	42	55	223
25-75%	87	148	336	218	789
75-90%	7	22	43	12	84
90-100%	18	9	12	1	40
Total	361	402	446	292	1501

Note: For schools where 500 or more students were tested, the passing rates ranged from 5.5% to 98.8%; for schools where more than 100 to 499 students were tested, the passing rates ranged from 1.7% to 96.5%.

Student Questionnaire

At the end of each test, students completed a brief questionnaire on their reactions to the test and their plans for high school and beyond. We examined the responses to these questions separately for students who did or did not pass each of the two tests. Tables 3.11–3.17 show the results.

TABLE 3.11 How did you prepare for this test?

Response	Failed ELA	Passed ELA	Failed Math	Passed Math
A. A teacher or counselor told me about the purpose and importance of the test	23.4%	34.7%	28.5%	30.6%
B. I practiced on a sample test	6.2%	6.4%	7.6%	7.1%
C. A teacher spent time in class getting me ready to take the test.	15.4%	19.5%	19.1%	16.3%
D. I did not do anything to prepare for this test.	22.1%	30.5%	33.0%	44.5%
No Response	32.9%	8.9%	11.8%	1.5%

TABLE 3.12 How important is this test to you?

Response	Failed ELA	Passed ELA	Failed Math	Passed Math
A. Very important	46.1%	52.8%	59.6%	52.7%
B. Somewhat important.	14.0%	22.5%	20.3%	30.6%
C. Not Important	3.0%	3.6%	3.3%	5.5%
No Response	37.0%	21.1%	16.8%	11.2%

TABLE 3.13 Do you think you will graduate from high school?

Response	Failed ELA	Passed ELA	Failed Math	Passed Math
A. Yes	43.7%	73.3%	63.7%	84.5%
B. No	2.1%	0.5%	2.1%	0.6%
C. Not sure	17.1%	5.0%	17.4%	3.7%
No Response	37.0%	21.2%	16.9%	11.2%

TABLE 3.14 Will it be harder to graduate if you have to pass a test like this?

Response	Failed ELA	Passed ELA	Failed Math	Passed Math
A. Yes, a lot harder	27.2%	16.4%	34.5%	13.4%
B. Somewhat harder	19.2%	33.8%	31.3%	38.1%
C. Not much harder at all	6.7%	20.9%	8.7%	31.1%
D. I really don't know	9.8%	7.6%	8.6%	6.1%
No Response	37.1%	21.3%	17.0%	11.3%

TABLE 3.15 What do you think you will do after high school?

Response	Failed ELA	Passed ELA	Failed Math	Passed Math
A. I will join the military	6.5%	3.4%	7.1%	6.6%
B. I will go to community college	10.1%	7.6%	12.4%	6.6%
C. I will go to a four-year college or university	25.3%	55.1%	38.5%	64.3%
D. I will go to Vocational/Technical/Trade School	2.1%	1.6%	2.3%	1.6%
E. I will work full-time	4.8%	1.2%	4.6%	1.0%
F. I really don't know what I will do after high school	13.4%	8.7%	17.2%	10.6%
No Response	37.8%	22.3%	18.0%	12.4%

TABLE 3.16 How sure are you about what you will do after high school?

Response	Failed ELA	Passed ELA	Failed Math	Passed Math
A. Very sure	25.8%	36.5%	34.4%	40.6%
B. Somewhat sure	25.8%	34.2%	35.4%	38.1%
C. Not sure at all	11.2%	8.1%	13.3%	10.1%
No Response	37.2%	21.2%	17.0%	11.3%

TABLE 3.17 How well did you do on this test?

Response	Failed ELA	Passed ELA	Failed Math	Passed Math
A. I did as well as I could.	42.5%	63.5%	53.5%	66.6%
B. I did not do as well as I could have, because	19.6%	14.7%	28.9%	21.7%
A. I was too nervous to do as well as I could.	12.3%	6.8%	10.6%	4.4%
B. I was not motivated to do well.	5.9%	4.5%	6.0%	4.8%
C. I did not have time to do as well as I could.	5.5%	4.1%	2.7%	1.4%
D. There were questions on this test that covered topics I was never taught.	6.8%	3.8%	18.7%	11.5%
E. There were questions on this test that covered topics I was taught, but I did not remember how to answer them.	6.2%	4.0%	14.6%	12.9%
F. There were other reasons why I did not do as well as I could have.	11.2%	10.4%	10.3%	8.1%
No reason checked.	0.1%	0.1%	0.2%	0.2%
No Response	37.9%	21.8%	17.6%	11.6%

Test Score Accuracy

A key question is how accurately students were classified as having achieved or failed to achieve the passing standard. We fit a statistical model based on item response theory to estimate how often students at each score level would be correctly classified. In our June 2000 report, we constructed a number of “pseudo-forms” and then estimated classification accuracy for each form. The procedure used here was the same except that we used data on the actual test form.

Data from the March administration were used to estimate item parameters for each test question. These parameters provide a prediction function giving the probability of a correct response (or of each score level for the essay questions) as a function of the student’s standing on an unobserved achievement scale. We selected 100 points along the IRT ability scale, corresponding to percentile points, so that each point represented one percent of the student population. For each point, we computed the probability of each possible pattern of correct and incorrect answers and, for the ELA test, each possible pattern of essay question score levels³. Each pattern corresponded to a specific number correct score. For mathematics, the number correct score was just the number of correct answers. For ELA, the number correct score was the weighted average of the number of multiple-choice questions answered correctly and the sum of scores on the two essay questions. By observing the probability of different patterns of number correct scores, we can estimate how much the student’s observed score from a single testing will differ from his/her “true” score (the average of scores from a large number of parallel administrations). Specifically, for each “true” score level, we estimated the proportion of time a student at that level would obtain an observed score that was above or below the passing level. We then compared these proportions to the student’s classification based on his/her “true” score to determine the percent of time the

³ Under the statistical models used, the “conditional” probabilities of correct answers to different test items are independent. This means that the probability that a student *at a given ability level* passes two different items is the product of the passing probabilities for each of the individual items.

student would be correctly and incorrectly classified as passing the test. Table 3.18 summarizes the expected scores and error of measurement for students at different percentiles. The errors of measurement shown in this table, while interesting, are not the most important indicators of accuracy for a test used to classify students as above or below a given level. We were interested, instead, in a measure of the accuracy of the classification decisions. While several researchers have worked on indicators of classification accuracy, we have developed our own approach to characterizing the accuracy of a test used for classification decisions. The basic concept is to divide the score scale into four regions. The passing level divides the upper and lower two regions. Students at levels 1 and 2 have true scores that are below the passing level and students at levels 3 and 4 have true scores above the passing level. The dividing point between levels 1 and 2 is the point at which a student will have an exactly 10 percent chance of passing from a particular testing session. Students in level 1 are below the passing point and have a greater than 90 percent chance of being accurately classified as being below passing. Students at level 2 are near enough to the passing point to have a significant chance of misclassification, given the accuracy of the test. Similarly, the point at which a student has exactly a 90 percent chance of passing divides levels 3 and 4. Students at level 3 are also near enough to the passing point to have a significant chance of misclassification, while level 4 students are fairly certain to be correctly classified as passing.

TABLE 3.18 Error of Measurement

Percentile	ELA			Mathematics		
	Expected Raw Score	Std. Error of Measurement	Probability of Passing (Pct.)	Expected Raw Score	Std. Error of Measurement	Probability of Passing (Pct.)
1	16.28	4.12	0.0	18.63	3.66	0.0
10	35.42	6.94	0.1	23.42	3.93	0.0
20	46.14	6.03	8.1	27.83	4.08	0.0
30	53.27	5.16	48.5	32.21	4.16	0.4
40	58.90	4.54	86.9	36.74	4.17	5.3
50	63.67	4.10	98.2	41.35	4.14	30.2
60	68.16	3.74	99.9	46.37	4.04	76.2
70	72.27	3.41	100.0	51.55	3.89	98.0
80	76.46	3.04	100.0	57.50	3.64	100.0
90	80.82	2.58	100.0	64.63	3.20	100.0
99	86.94	1.57	100.0	76.95	1.62	100.0

Levels 2 and 3 constitute a “zone of uncertainty” where correct classification is at risk. As shown in Table 3.19 below, between 37% and 38% of the students whose true score was at level 2 actually passed the exam. Similarly, between 27% and 30% of the students in zone 3 failed to pass. Outside this zone of uncertainty, the rate of correct classification ranges from 96% (level 1) to 99% (level 4).

TABLE 3.19 Classification Error

True Achievement Level	ELA			Mathematics		
	Raw Score Range	Pct. in Range	Pct. Passing	Raw Score Range	Pct. in Range	Pct. Passing
1, Well Below Cut	00.0-46.5	19.9	3.7	00.0-42.6	51.4	2.4
2, Slightly Below Cut	46.5-54.0	10.4	37.0	42.6-44.0	2.0	37.8
3, Slightly Above Cut	54.0-59.7	10.5	70.7	44.0-48.6	9.8	73.1
4, Well Above Cut	59.7-90.0	59.2	98.8	48.6-80.0	36.7	98.8

The classification accuracy of a test may be characterized by the narrowness of the zone of uncertainty (levels 2 and 3) and by the proportion of examinees that falls outside the uncertain range. For ELA, the zone of uncertainty ranged from 46.5 to 59.7 raw score units, corresponding to 51.7% to 66.3% of the possible 90 points. Of these students, 79% fall outside the zone of uncertainty. For mathematics, the zone of uncertainty is narrower, ranging from 53.3% to 60.8% of the 80 possible points. In addition, 88% of the students were outside the zone of uncertainty on the mathematics test.

At their December 2000 meeting, the SBE approved revised test specifications that included fewer questions for each of the two exams. Both tests were shortened relative to the original specifications, from about 100 multiple choice questions down to 80 to 82 questions. The result was inevitably some loss in the accuracy of the test scores and the precision with which students are classified as above or below the passing standard. The accuracy of the ELA test is further affected by the relatively large weight given to the two essay questions in comparison to the multiple-choice. Nonetheless, both tests appear to be performing reasonably well. Between 80% and 88% of the students are unambiguously classified as being above or below the passing standard. For the remaining students, their true achievement is quite near the passing standard. The consequences of passing a modest number of students who are only slightly below the standard while requiring a modest number who are barely above the standard to retest would not appear to be serious.

Two qualifications are in order. First, there are no hard standards for classification accuracy. The tendency has been to fall back on traditional estimates of test reliability based on the ratio of measurement error to total score variance across the whole range of the test. Second, the estimates of the proportion of students whose true achievement falls in each range and the percent passing within each range are based on assumptions underlying particular statistical models⁴.

⁴ We used the 3-parameter logistic model for the multiple-choice questions to accurately model the effects of guessing. We used an 8-level partial credit model (Muraki, 1992) for each essay question to model the 8 possible scores a student might receive based on the combination of two independent ratings.